Root-Word Analysis of Turkish Emotional Language

Ozan Cakmak[1], Abe Kazemzadeh[2], Dogan Can[2], Serdar Yildirim[1], and Shrikanth Narayanan[3]

Department of Computer Engineering, Mustafa Kemal University[1]

Department of Computer Science, University of Southern California[2]

Department of Electrical Engineering, University of Southern California[3]

ozancakmak@mku.edu.tr, kazemzad@usc.edu, dogancan@usc.edu, serdar@mku.edu.tr,shri@sipi.usc.edu

Abstract

This paper describes a model for the perceived emotion of Turkish sentences based on the emotions associated with the constituent words. In our model, each emotion is mapped to a point in the continuous space defined by three emotional attributes: valence, activation, and dominance. We collected a large data set through two independent surveys: a word-level survey that prompted users with emotional words and asked them to assign each word a continuous emotional interval, and a sentence-level survey that prompted users with emotional sentences collected from 31 children's books and asked them to rate each sentence on a discrete emotional scale. The word-level survey was aimed at creating a core affective lexicon for Turkish. It is difficult to build a comprehensive affective lexicon for Turkish due to its very productive morphology that generates a very large vocabulary. We deal with the sparsity issues caused by the large word vocabulary by analyzing the emotional content of word roots. Our experimental results indicate that there is a strong correlation between the emotions attributed to Turkish word roots and the Turkish sentences.

Keywords: affective computing, emotion recognition, sentiment analysis, emotion analysis and annotation

1. Introduction

Automatically analyzing the emotional content of language has become increasingly important for applications that deal with natural language. For instance, the tasks of opinion mining and affective computing (Picard, 1997) are receiving a lot of attention in the fields of Natural Language Processing, Fuzzy Logic Systems such as an interval type-2 (Kazemzadeh et al., 2008), and Human Computer Interaction (Fragopanagos et al., 2005). Despite the progress of previous works (Jang and Shin,2010) in the field, there has been relatively less progress in non-English languages. The study of other languages within affective computing offers new technical and scientific challenges. We believe that our study opens new perspectives and brings about new methods that can increase the applicability of natural language affective computing to more diverse languages. In this study, we analyze Turkish(Katzner, 2002), which is an agglutinative language, which means that new words can be formed from existing words by a rich set of affixes (Oflazer, 1994).

The unique characteristics of Turkish present various challenges to current approaches to emotional analysis by natural language processing because the agglutinative word formation process create many unique words. In our study, we observed that there is a strong correlation between the emotions attributed to word roots, which are the core forms of words, and the emotion of sentences when negations, derivations, and inflections are accounted for. We measured correlation empirically from annotations of words and sentences in terms of valence, activation, and dominance (Russell and Mehrabian, 1977). A perennial challenge in affective computing research is the availability of suitable data resources. We have created a novel corpus of Turkish text from children's books richly annotated with affective information using crowd sourcing techniques. This corpus is large by the standards of other comparable emotional corpora and is one of the first emotionally labeled corpora for Turkish.

The reason why we chose this approach is that a single root might produce many different word forms in an agglutinative language like Turkish. Our hypothesis is that it is the emotion of the constituent word roots that determines and identifies the perceived emotion of sentences. However, such an analysis is not so simple because the effects of affixes like negation, which change the meaning of the roots, present theoretical contradictions to this general hypothesis. These affixes, which can potentially change the meaning of the root words, must be treated differently from the set of other affixes. Broadly, the affixation process can be seen in terms of phonological rules (e.g., vowel harmony, where vowel characteristics become assimilated in the neighboring vowels), derivational rules (e.g. grammatical recategorizations such as nominalization, which derives a noun from a verb), and inflectional rules (e.g. verb tense and noun pluralization).

2. Methodology

In our study, we analyze the data at both the word and the sentence level. Our sentence-level data comes from 31 children's books such as world classic novels, fairy tales, stories of heroism, romance, etc. Children's books were chosen to make up the corpus since these books contain a wide array of easily identifiable emotions. This corpus consists of 83,120 sentences. It contains 1,045,297 words, 110,695 of which are unique. The high number of unique words reflects the agglutinative nature of Turkish. The corpus is annotated at the sentence-level with one of the seven emotion categories (Angry, Happy, Sad, Disgusted,

Neutral, Surprised, and Fear) as well as valence, activation and dominance values, which are what we focus in this study. Valence measures whether the emotion is negative (unpleasant) or positive (pleasurable). Activation measures how strong the emotion is: dispassionate (calm) or passionate (excited). Dominance measures how assertive the emotion is: submissive (retreating) or dominant (aggressive). In our corpus, one point on the scale from 1-9 is used to represent these emotion characteristics. The corpus was distributed to 31 college students, who sequentially annotated the sentences with emotion category labels and valence, activation, and dominance values. To deal with the agglutinative word constructions, we extracted word roots with the Zemberek[1] Library, which is an open source, general purpose Natural Language Processing library for Turkish.

Let's take a look at example emotional words(Table 1-3):

Turkish	Öfkeli
English	Furious
Gloss	(root:öfke/fury) + (affix:-li/-ous, adjectival derivation)

Table 1: Example emotion word "furious"

Turkish	Hevesli
English	Zealous
Gloss	(root:heves/zeal) + (affix:-li/-ous, adjectival derivation)

Table 2: Example emotion word "zealous"

Turkish	Dertli
English	Sorrowful
Gloss	(root:dert/sorrow) + (affix:-li/-ful, adjectival derivation)

Table 3: Example emotion word "sorrowful"

After decomposing the words into root and affixes using the Zemberek Library(Akin et al., 2008), our corpus had 10,018 unique word roots. The word root can be seen as the basic component of a word's meaning after removing phonological, inflectional, and derivational effects. Our hypothesis is that the level of the root words is the best way to analyze Turkish sentences emotionally. However, stripping the words to their roots ignores critical derivations like negation. To measure the effects of these critical affixes, we performed two experiments: one, which removed these derivations, and another, which left them intact.

[1] http://code.google.com/p/zemberek/

English	**Turkish**
enthusiasm	Şevk
terrible	Berbat
courage	Cesaret
mad	Çılgın
tired	Yorgun
calm	Sakin
hopeful	Ümitli
interested	Ilgili
surprised	Şaşkın
boredom	Sıkıntı
sadness	Üzüntü
expectation	beklenti
worried	endişeli
lucky	Şanslı
happy	Mutlu
amusement	Eğlence
assidious	gayretli
confidence	Itimat
willing	Istekli
lucky	Şanslı
mercy	merhamet
patient	Sabırlı
love	Sevgi
joyful	sevinçli
admiration	Hayran
fear	Korku
frustration	Hüsran
arrogant	Kibirli
depression	depresyon
nervous	Sinirli
pleasure	memnuniyet
sympathy	sempati
proud	Gururlu
restful	huzurlu
excited	heyecanlı
heroism	kahramanlık
honorable	Onurlu

Table 4 : Some words from 197 Emotion Words

To measure the word-level emotion characteristics, we conducted a survey[2] of approximately 40 people who were presented with 197 emotion words (Table 4) and asked to rate these on valence, activation, and dominance scales. These words came from the EMO20Q Project(Kazemzadeh et al., 2011), which uses the emotion twenty questions game as a way to observe the human intuition about emotions. We translated 171 words from

[2] http://sail.usc.edu/~kazemzad/fuzzyEmotionEvaluation/turkish/turkish_experiment1.cgi

this project to Turkish and additionally added 26 synonyms. The emotional rating scales for this survey are different from the corpus annotation task in that two points are used for the scale, one to present the lower bound of a range of possible values and the other for the upper bound, which allows for measurement of intra-subject uncertainty. Also, the survey's scales ranged from 0 to 100. The survey consisted of four sessions per subject wherein each subject was presented with thirty-five words chosen randomly from the set of 197 words. This resulted in each of the 197 words being rated approximately 30 times. To compare the single-point scale of the sentence-level annotations to the double-point (upper and lower) scale of the word-level annotations, we converted the (upper-point, lower-point) representation into the (midpoint, radius) form.

Of the 197 emotion word roots from the survey, twenty-four did not occur in the corpus. As a result, the total count of word roots for the survey and the corpus is 173. In addition, in both the corpus and the survey, 99 emotion words were carefully chosen without possible derivational negations (the affixes -siz, sız, -suz and -süz), which can potentially change the emotion of word root. We separately analyze this subset and its complement.

Let's take a look at these examples (Table 5-6):

Turkish	ilgi -li
English	interested
Gloss	(root: ilgi/interest) + (affix: -li/-ed, adjectival derivation)
ANTONYM	
Turkish	ilgi-siz
English	un-interest-ed
Gloss	(root: ilgi/interest) + (affix: -siz/un-...-ed, negative adjectival derivation)

Table 5: Example emotion words "interested" and "uninterested"

Turkish	ümit-li
English	Hopeful
Gloss	(root: ümit/hope) + (affix: -li/-ful, adjectival derivation)
ANTONYM	
Turkish	ümit -siz
English	hope-less
Gloss	(root: ümit /hope) + (affix: -siz/-less, negative adjectival derivation)

Table 6: Example emotion words "hopeful" and "hopeless"

Although these words contain the same root, the derivational suffixes completely change the emotional connotation, in this case valence. To see the effects of these affixes, we performed correlation analysis both with and without these affixes.

3. Results

The 173 emotion word roots described above were identified in the corpus and the average sentence valence, activation, and dominance were calculated for each word root. Then, we compare, using correlation, these sentence-level averages with the word-level average valence, activation, and dominance values from the surveys. We found moderately high correlation between the word and the sentence-level valence (rho=0.55) and lower correlation for activation and dominance (rho=0.29 and rho=0.20, respectively). Then we repeated the correlation analysis on a subset of words having no negation present (99 words) and another subset having negation affixes (74 words).

Correlation	Valence
All words(173)	0.55
Words without negation and derivational affixes(99)	0.65
Words with negation and derivational affixes(74)	0.47

Table 7: Correlation Results for Valence.

Correlation	Activation
All words(173)	0.29
Words without negation and derivational affixes(99)	0.31
Words with negation and derivational affixes(74)	0.23

Table 8: Correlation Results for Activation.

Correlation	Dominance
All words(173)	0.20
Words without negation and derivational affixes(99)	0.24
Words with negation and derivational affixes(74)	0.10

Table 9: Correlation Results for Dominance.

We found that the subset without negation had a stronger correlation than the mixed set and the set containing negation affixes, and furthermore, that the set with negations had the lowest correlation values. This correlation of the averages of valence, activation and dominance values between the corpus and the survey indicates that perceived emotion of sentences is highly correlated with the chosen specific emotion words (Table

7-9). The stronger correlation in the valence dimension indicates that valence is the most strongly lexicalized emotional attribute.

4. Conclusion

In this paper, we verified that the emotions attributed to Turkish word roots are highly correlated with the emotion of Turkish sentences. We found that the emotional characteristics of sentences in terms of valence, activation, and dominance are significantly correlated with the emotional characteristics of the constituent words, when the words are decomposed into roots, and that moreover taking into account the exception of negation affixes makes this correlation stronger. This shows that negation affixes can significantly modify the emotion of words and sentences.

In our study, we measure the effects of this factor so that it can be taken into consideration in future studies. This approach of root analysis can be applied to various applications for extracting important emotions on the Internet, mobile phones or human computer interaction applications to make social networks for people who have similar opinions. Although English is not an agglutinative language, it also contains affixes that modify root words, so our results may be applied to non-agglutinative languages as well.

We plan to confirm the results of this paper by experiments on the survey and the corpus, which will be analyzed in more detail to consider negations, derivational affixes and inflectional suffixes. In addition to studying the relation of the word and sentence-level emotional scales, we also plan to examine the inter- and intra-subject variability. Inter-subject variability can be analyzed in terms of agreement between subjects and intra-subject variability can be seen in coherent behavior on repeated stimuli and by leveraging the upper and lower-points of the word-level surveys, which were designed for fuzzy logical analysis of emotional meaning.

Also, we plan to study the categorical labels of the sentence-level corpus. We plan to share this corpus, which is large by the standards of other comparable emotional corpora and one of the first emotionally labeled corpora for Turkish.

5. Acknowledgements

We are very grateful to all the annotators. This research was developed in the context of FLS (Fuzzy Logic System) for Turkish Language.

6. References

R. Picard. 1997. Affective Computing, MIT Press.

A. Kazemzadeh, S. Lee, and S. Narayanan. 2008. An interval type-2 fuzzy logic system to translate between emotion-related vocabularies, in Proceedings of Interspeech, (Brisbane, Australia).

N. Fragopanagos, J. G. Taylor. 2005. Emotion in human-computer interaction, Neural Networks, Volume 18, Issue 4, Pages 389-405

H. Jang and H. Shin. 2010. Language-Specific Sentiment Analysis in Morphologically Rich Languages, Proceedings of Coling, (Beijing)

Kenneth Katzner. 2002. The Languages of the World, 3rd Ed., Routledge

Oflazer Kemal. 1994. Two-level Description of Turkish Morphology, Literary and Linguistic Computing, vol. 9, No:2.

J. A. Russell and A. Mehrabian. 1977. Evidence for a three-factor theory of emotions. Journal of Research in Personality, vol. 11, pp. 273-294.

Ahmet Afsin Akin, Mehmet Dundar Akin.2007. Zemberek, an open source NLP framework for Turkish Languages.

A.Kazemdadeh, P.G. Georgiou, S. Lee, and S.Narayanan. 2011. Emotion questions: Toward a crowd-sourced theory of emotions, in Proceedings of AC'II,2011